Is an account of the illusion of consciousness conceivable?

Lars Š. Laichter[1] —— April 29, 2019

1 Introduction

Explaining the nature of conscious experiences is considered to pose a unique epistemic challenge. This challenge arises due to an inconsistency between our concepts of conscious experiences and the physicalist picture of the world, as conscious experience is often claimed to be non-physical. Illusionism is a theory which reconciles this inconsistency by asserting that conscious experience is not real and that it is only a product of an illusion. In this paper, I discuss a noteworthy problem for the illusionist theory known as the illusion meta-problem. The illusion meta-problem highlights the difficulty of conceiving of the proposed illusion of consciousness as true. Given this difficulty, I ask whether an account of the illusion of consciousness is conceivable in principle. My thesis is that a functional account of the illusion of consciousness is conceivable if the account meets four criteria: the definitional criterion (an explicit definition of illusion), the functional criterion (an account of illusion in terms of functionalism), the framework criterion (a generalisation in terms of representations or other frameworks of the mind), and the empirical criterion (empirically supported commonalities between the illusion of consciousness and other instances of illusions). To argue this thesis, I establish the necessity of each one of the proposed criteria in regards to the conceivability of the illusion of consciousness. In addition, I provide an outline of possible strategies to meet each one of the proposed criteria. My four criteria constrain the conceivability of the illusionist account, thereby paving a possible way for a functional account of the illusion of consciousness.

[1] Email: lars.laichter@gmail.com

2 Background

I begin by introducing the non-physicalist[2] notion of conscious experience. In general, the concept of conscious experience is said to be equivalent to the notion that there is something “it is like” to be an organism from a first-person perspective (Nagel 1974, p. 438). Conscious experience is also sometimes referred to as phenomenal consciousness or qualia. I will use the term conscious experience as equivalent to these terms throughout this paper. Some examples of conscious experiences include experiencing the redness of red, tasting the taste of coffee, and experiencing the feeling of pain. Nagel famously claims that one can never know what “it is like” to be a bat from a third-person perspective. However, Nagel’s notion that “it is like” something to be an organism does not directly highlight the trouble of explaining conscious experience in physical terms. The trouble of explaining conscious experience in physical terms is sometimes said to arise due to unique properties of consciousness. These properties include ineffability, intrinsicality, privacy, and lack of direct or immediate apprehensibility (Dennett 1988, p. 3). There are two popular ways of introducing the concept of conscious experience: either by appealing to its presence in its seeming absence or by appealing to its possible absence in its seeming presence.

First, one might appeal to the presence of conscious experience in its seeming absence. The most popular example of such an approach is the thought experiment about Mary (Jackson 1982, p. 130). Mary lives in a black and white room. She is fascinated by colour, although, she has never seen it. To satisfy her curiosity, she learns all physical information about colour. This includes everything from the physics of light to the biochemical processing of light rays in the nervous system. However, one day she is able to leave the room, and she sees colour for the first time. Jackson argues that in that very moment she learns something new about the world and the visual experience of it. She learns “what it is like” to have the conscious experience of seeing colour. Since she already knew all the physical information, it must be that her knowledge was incomplete and she must have learned something which cannot be learned only from physical information. Jackson concludes that if she could not have learned about conscious experience from physical information, then it must be that conscious experience cannot be described

[2] I use the notion “non-physicalist” as one opposing the physicalist hypothesis which posits that everything supervenes on the physical.

in physical terms. The argument introduced by the thought experiment about Mary can be put like this:

> (1) Mary has all the physical information concerning colour, including human colour vision, before she leaves the black & white room.
> (2) There is some information that Mary does not know about colour vision that she gains once she leaves the black & white room.
>
> ---
>
> (3) Not all information is physical information.

The argument featured in the thought experiment about Mary has been contested in various ways, but the thought experiment remains to be one of the prime examples of contesting physicalism about the mind.

Second, one might appeal to the possible absence in its seeming presence. The most well-known example of such an approach is the example of philosophical zombies popularised by Chalmers (1996, p. 95). Philosophical zombies are physical duplicates of humans which are physically and functionally identical, but which lack conscious experience. A philosophical zombie is not meant to refer to zombies in the sense one might be familiar with from popular culture. It is rather meant to refer to a copy of a human that can behave indistinguishably from normal humans, with the difference of not having any conscious experience. Chalmers' argument for the existence of philosophical zombies is proceeding from conceivability to possibility. In other words, if philosophical zombies are conceivable, then such zombies are possible. Chalmers uses this entailment to propose an argument against physicalism. The argument can be roughly put as follows:

> (1) Philosophical zombies are conceivable.
> (2) If philosophical zombies are conceivable, philosophical zombies are possible.
> (3) If philosophical zombies are possible, physicalism is false.
>
> ---
>
> (4) Physicalism is false.

The argument is dependent primarily on the entailment of possibility by conceivability, as well as on the assumption whether philosophical zombies are in fact conceivable. Nevertheless, the thought experiment is another prime example of contesting physicalism about the mind.

Both of these examples highlight the problem of explaining the nature of conscious experience. The problem of explaining conscious experience is also known as the hard problem of consciousness (Chalmers 1995, p. 200). The hard problem of consciousness is often used to distinguish the problem of explaining conscious experience from other "easier" problems–also known as the easy problems of consciousness. The easy problems are problems which are explainable in physicalist terms and, therefore, do not give rise to an inconsistency between conscious experience and physicalism. An explanation of an easy problem might take various forms, including a functional explanation (one concerned with a particular function of an apparatus on its own or within a system) or an account of a neural mechanism in physical terms. Examples of easy problems include reportability of mental states, access to internal states, the focus of attention, deliberate control of behaviour, or the difference between wakefulness and sleep. It is not the case that science necessarily has explanations of all the "easy" problems of consciousness, but there exists a method via which science can conceivably obtain such explanations. In contrast to the easy problems, the hard problem is supposed to lack such a method due to its non-physical properties outlined by the above-introduced though experiments. In short, the hard problem of consciousness is the problem of explaining the nature of conscious experience. A possible solution to the hard problem is the subject of the following paper.

Illusionism

Illusionism is a theory which aims to solve the hard problem by claiming that conscious experiences are a product of an illusion and, therefore, they do not exist. If conscious experiences do not exist, there is no hard problem of consciousness. Thus, by establishing that conscious experiences are illusions, the illusionist theory has the capacity to practically "explain away" the hard problem. The argument can be roughly put as:

> (1) The hard problem aims to explain the existence of conscious experiences.
> (2) All conscious experiences are a product of an illusion.
> (3) If conscious experiences are a product of an illusion, they do not exist.

(4) If what the hard problem aims to explain does not exist, there is no hard problem.

---

(5) There is no hard problem.

Although illusionism remains a minority view, the view has a long-standing history and has recently enjoyed a surge in popularity, partly due to the work by Frankish. Frankish formulates illusionism in terms of the illusion problem, which he claims to be the core problem illusionists need to solve (2016, p. 12). In his account, the illusion problem replaces the hard problem. The illusion problem asks why is it the case that conscious experience seems to exist and what is the particular physical mechanism that gives rise to the illusion of consciousness. Although there are various forms of illusionism, I will focus on what is referred to as "strong illusionism." According to the "strong" illusionist view, conscious experience in its entirety is a product of an illusion and not just some of its parts. This particular version maintains the capacity to solve the hard problem, as it commits to the premise that *all* rather than just *some* conscious experiences are a product of an illusion. The illusionist theory has been challenged on a number of its premises. Perhaps the strongest objection comes in the form of the illusion meta-problem.

The illusion meta-problem

The illusion meta-problem raises an objection about the conceivability of the illusion of consciousness which is the central question of this paper. The illusion meta-problem was originally proposed by Kammerer (2016, p. 125). The illusion meta-problem asks for an explanation of why the illusion of consciousness appears more convincing than other illusions. Kammerer claims "the fact that the illusionist hypothesis strikes many of us as wildly implausible, deeply puzzling, almost absurd, and so on, is a fact about the particular way in which we are subject to the illusion of consciousness" (2018, p. 7). This can be summed as the notion that the illusion of consciousness is somehow more convincing than other illusions. By convincing, I mean that it appears particularly hard to conceive of the proposition that our conscious experiences are illusory as true. For example, I might be able to conceive of the proposition that my hands are only illusory as true (Kammerer 2019, p. 125). It might be the case that there is just a very elaborated implant in my brain which precisely simulates all necessary sensory inputs while I actually interact with the world with bionic limbs. I might not know about this implant since I might have had it since birth. It would not strike me as particularly difficult to conceive of such a scenario. Nonetheless, the proposition of my conscious experience being an illusion appears much more difficult to conceive of. The illusion meta-problem claims that it is difficult to conceive of the illusion of consciousness. However, it does not establish whether a possible solution to the illusion problem is conceivable at all. It is precisely the question whether the illusion of consciousness is conceivable in principle that is a subject of the following discussion.

Conceivability

The question confronted in this paper is the question of conceivability of the illusion of consciousness. Let us begin by specifying the the term of conceivability I use the term "conceivability" to refer loosely to the capacity to form a consistent conception of a particular concept. Thus, in the case of the following discussion, when I ask whether we can conceive of the illusion of consciousness, I ask whether we can form a consistent conception of the concept of the illusion of consciousness. I am not asking about the capacity of a particular person and their capacity to conceive of a concept, but rather about conceivability in a universal sense. Conceivability is sometimes established in terms of an agent with an ideal reasoning (Chalmers 2002). To establish the notion of conceivability in this case, the agent would have to be ideally supplemented with the identical condition of being subjected to the illusion of consciousness. Hence, to conceive of an account of the illusion of consciousness means to form a consistent conception of the concept of the illusion of consciousness while the agent is subjected to the illusion of consciousness.

When examining the conceivability of a concept, there are presumably two outcomes: either the concept is conceivable or it is inconceivable. An example of a conceivable concept is the mechanism by which keystrokes are translated to letter on the screen of my computer. There is a procedure which someone has conceived of to implement the mechanism. Although I do not know the specific of the mechanism, its consistency could be verified using a series of computer science and mechanical engineering concepts. Therefore, it appears for such a mechanism to be conceivable. On the other hand, an example of an inconceivable concept might be a general method to decide whether any particular proposition is true or false. Such a method does does not exist, as proven by Turing (1937). Since there is not such a method, one cannot conceive of such a method under the logical constraints of the problem.

However, one might ask what makes the first example conceivable. Ultimately, the claim concerning conceivability a mechanism implemented in my computer is an epistemic claim which is a subject to relevant criteria. On one hand, I could claim that there are little men that are painting letters on my screen according to my keystrokes. On the other hand, I could require an infinitely detailed explanation of the molecular dynamics the hardware of my computer. Neither of these accounts constitutes a reasonable conception of my keyboard mechanism. To obtain a reasonable conception, one would constrain the account via a set of relevant criteria. In the case of the keyboard, it might be the appropriate mapping of keystrokes to characters on the screen or the basic rules of hardware construction. In the following discussion, I attempt to establish the necessary criteria for an account of the illusion of consciousness.

3 The definitional criterion

The four following sections describe the necessary criteria for the conceivability of a functional account of the illusion of consciousness. Each one of these sections is divided into two major parts: (1) an argument for the necessity of the particular criterion, and (2) a possible solution to that particular criterion. I begin with the definitional criterion which corresponds to the claim that it is necessary for a functional account of the illusion of consciousness to entail an explicit definition of what constitutes an illusion.

Arguments for necessity usually take on the following form: If P is necessary for Q, then Q cannot be true unless P is true. To establish the definitional criterion, I argue that a definition of illusion must be conceivable for an account of the illusion of consciousness to be conceivable. The argument for the necessity of the definitional criterion can be put as:

(1) A definition of illusion is conceivable.
(2) An account of the illusion of consciousness is conceivable.
(3) If *(2)* cannot be true unless *(1)* is true then *(1)* is necessary

for *(2)*.
(4) *(2)* cannot be true unless *(1)* is true.

---

(5) *(1)* is necessary for *(2)*.

The particular premise in question when arguing for the necessity of this criterion is the premise (4). The reason why the premise (4) is true is because the illusion of consciousness is arguably sharing some features with other illusions. If the illusion of consciousness shares features with other illusions, it must be necessary for someone who is to conceive of an account of the illusion of consciousness to be also able to conceive of a definition of illusion in general. Let us suppose the contrary. In such an instance, there would be little reason for one to believe that conscious experience could be conceivably accounted for in terms of illusion, as illusion itself would not be conceivable. Therefore, a definition of illusion must be conceivable for an account of the illusion of consciousness to be conceivable.

Considering premise (1), there is little chance that a definition of illusion is not conceivable Nonetheless, the illusionist discourse relies mostly on an implicit notion of illusion. There are various senses of the word illusion. For example, one could say that "I have no illusions about the quality of my philosophical writing" or "I had the illusion of my hands disappearing." The first example refers to a particular belief while the second example refers more to a deceptive appearance. There are possibly other senses in which the word is used that I do not include here. One of the strong objections to the implicit use of illusion in the illusionist discourse is the claim that hallucinations are a better suited to explain the seeming appearance of the illusion of consciousness. Chrisley and Sloman claim that the term 'hallucination' is a better suited term to explain the illusion of consciousness (2016, p. 2). Their claim is primarily supported by a particular distinction in the meaning of the words illusion and hallucination. The distinction rests on the notion that illusion refers to instances when "we seem to see objects as other than they are" and hallucination refers to instance when we "seem to see objects that are not there" (Fish 2009, p. 4). As they claim, there is no object to be perceived in the case of consciousness, as it is the case in the general understanding of hallucination. This conflict concerning the use of the notion of illusion within the illusionist discourse highlights both the need for an explicit definition and a definition that is broad enough to encompass the possibly more intricate illusion of consciousness, as well as other instances of illusions. Thereby, in the subsequent section I provide an explicit definition of illusion which shows that a definition of illusion is conceivable and addresses the objection by Chrisley and Sloman.

A definition of illusion

In this section, I provide an explicit definition of illusion which shows that a definition of illusion is conceivable. In addition, my definition addresses Chrisley's and Sloman's objection by including their notion hallucination in the definition of illusion. I introduce a definition of a super-illusion, as a possible definition that encompasses the illusion of consciousness, as well as other illusions. In addition, the definition of the super-illusion subsumes both the traditional notion of illusion and of hallucination. To define the super-illusion, I first define the traditional notion of an illusion as a property illusion and the traditional notion of hallucination as an existence illusion. I subsequently combine them to form a the definition of the super-illusion.

To introduce the proposed concept of super-illusion, I utilize a modified version of epistemic logic (Rescher 2005, p. 1). Epistemic logic is a branch of logic that seeks to formalize the logic of discourse about knowledge. I will make use of the resources of propositional and quantificational logic, supplemented by the machinery of quantified modal logic (Rescher 2005, p. 2). All of the following symbols will be used in the standard way:

$\neg, \wedge, \vee, \supset$ and $\equiv$ for traditional connectives
$\exists$ and $\forall$ for existential and universal quantification

I will also use some free logic to refer to the set of 'existing' objects (E!). Additionally, the following notational convention will be employed:

$x, y, z, ...$as variables to refer to agents
$o, m, n, ...$ as variables to refer to objects
Rxo ("x represents o")

For the sake of the the following analysis, I will also substitute the predicate K for alternative predicate I which stands for 'is under the illusion.' I also distinguish between I^E, I^P, and I^S which stand for *existence illusion*, *property illusion*, and *super-illusion* in the corresponding order. This predicate is meant to signify instances where some sort of illusion is present. The details of the predicate are specified in the subsequent definitions. I also use the predicate P and R^I. P corresponds to a predicate assigning an arbitrary denotation. R^I refers to instances of introspective representation, as apposed to simple representations R. When an agent introspectively represents something, it refers to an instance when an agent implements a causal structure which mirrors relevant causal properties of the object and the representation is available introspectively.

Existence illusion

Existence illusion is primarily aimed at accounting for instances when an agent represents an object as one which exists, although the object does not exist. This might specifically refer to some instances of hallucinations where we represent objects as existing, although they do not exist or are not present. Dreams are perhaps one of the most salient examples of such hallucinations. The definition for *existence illusion* can be phrased as:

> All agents are under the illusion of the existence of some objects if and only if they introspectively represent the object as existent while the object does not exist or if they represent the object as non-existent while the object exists.

We can also substitute the agents for ‘x’ and objects for ‘o’. In addition we can substitute ‘if and only if’ with $\equiv$ and establish that the definition holds for all agents and some object.

> $(\forall x)(\exists o)(x$ are under the illusion of existence of some $o \equiv x$ introspectively represents o as existent while o does not exist or if x represents o as non-existent while o exists).

For the formalization, I utilize a variable ‘x’ and ‘o’ where ‘x’ is an agent and ‘o’ is an object. I also use the predicate ‘I^E_ _’ for ‘agent _ are under the illusion of the existence of _’ and ‘R^I_ _’ for ‘_ introspectively represents _.’ In addition, I use E!_ to signify that that ‘_ belongs to a set of objects which exist.’ The definition for *existence illusion* can be defined as:

$$(\forall x)(\exists o)(I^E xo \equiv (R^I xE!o \wedge \neg E!o) \vee (R^I x\neg E!o \wedge E!o))$$

This definition asserts that an agent cannot simultaneously be under the illusion of an object as existing while knowing that the object exists and for the object in fact to exist or vice versa. One can also imagine such an assertion to be formulated for a proposition in terms of truth or falsity. However, as Chalmers points out “when I introspect my beliefs, they certainly do not seem physical, but they also do not seem nonphysical in the way that consciousness does” (2018, p. 23). I can assert that the colour of this page is green and represent it to myself as a proposition. However, it does not cause me to have any sort of illusion or hallucination whatsoever. Hence, to conserve the capacity of super-illusion of the consciousness, I will omit the capacity of beliefs in the form of propositions to be super-illusions. Although, we might say that “Mary is under the illusion that 1+1=3,” referring to Mary’s false belief, I do not think this refers to anything else than an extension of the use of the word ‘illusion.’ Such a use of the word ‘illusion’ does not seem to encompass what the illusion of consciousness seems to be. There is, however, an important element of illusion which I attempt to account via property illusion.

Property illusion

Property illusion is primarily aimed to account for instances when an agent represents an actually existing object to have properties which the object does not in fact have. This might specifically refer to some instances of perceptual where we represent objects as have certain properties, although they do not have them. The definition for *property illusion* can be phrased as:

> All agents are under the illusion of properties of some objects if and only if they introspectively represent the object as having a property P while the object does have a property P or if they represent the object as not having a property P while the object has a property P.

We can also substitute the agents for ‘x’ and objects for ‘o’. In addition we can substitute ‘if and only if’ with $\equiv$ and establish that the definition holds for all agents and all objects.

> $(\forall x)(\exists o)(x$ are under the illusion of property P of some $o \equiv x$ introspectively represents o as having a property P while o does have a property P or if x represents o as not having a property P while o has a property P).

For the formalization, I utilize a variable ‘x’ and ‘o’ where ‘x’ is an agent and ‘o’ is an object. I also use the predicate ‘I^P_ _’ for ‘agent _ is under the illusion of the property of _’ and ‘R^I_ _’ for ‘_ introspectively represents _.’ In addition, I use ‘P_’ for ‘_ has an arbitrary property.’ The definition for *property illusion* can be defined as:

$$(\forall x)(\exists o)(I^P xPo \equiv (R^I xPo \wedge \neg Po) \vee (R^I x\neg Po \wedge Po))$$

This definition asserts that an agent cannot simultaneously be under the illusion of an object having certain properties while the object indeed has these properties or vice versa. The definition neither necessitates that an agent is under an illusion which would include all properties of an object. Not is the assertion assuming that an agent does not know the correct property.

Super-illusion

There is no reason to believe that the illusion of consciousness cannot be composed of both existence illusion and property illusion. I refer to such a combination as the super-illusion. The super-illusion is a combination of existence illusion and property illusion. The conditional for *super-illusion* can be phrased as:

> All agents are under the illusion of properties or existence of some objects if and only if they introspectively represent the object as existent while the object does not exist or if they represent the object as non-existent while the object exists or if they introspectively represent the object as having a property P while the object does have a property P or if they represent the object as not having a property P while the object has a property P.

We can also substitute the agents for ‘x’ and objects for ‘o’. In addition we can substitute ‘if and only if’ with $\equiv$ and establish that the definition holds for all agents and all objects.

> $(\forall x)(\exists o)(x$ is under the illusion of existence or property P of some $o \equiv x$ introspectively represents some o as existent while o does not exist or if x represents o as non-existent while o exists or if x introspectively represents o as having a property P while o does have a property P or if x represents o as not having a property P while o has a property P).

For the formalization, I utilize a variable ‘x’ and ‘o’ where ‘x’ is an agent and ‘o’ is an object. I also use the predicate ‘I^S_ _’ for ‘agent _ is under the illusion _’ and ‘R^I_ _’ for ‘_ introspectively represents _.’ In addition, I use ‘P_’ for ‘_ has an arbitrary property’ and E!_ to signify that that ‘_ belongs to a set of objects which exist.’ The definition of *super-illusion* can be put as:

$$(\forall x)(\exists o)(I^S x[E!o \lor Po] \equiv [(R^I xE!o \land \neg E!o) \lor (R^I x \neg E!o \land E!o)] \lor [(R^I xPo \land \neg Po) \lor (R^I x \neg Po \land Po)])$$

Further work is required to establish the exact relations between property illusion and existence illusion. For example, it appears strange that one could be under an illusion of an object as having false properties while also representing the object as non-existent. An example of such an instance might be for example dreaming of an optic illusion in a dream. In such a case, an agent represents an object with properties the object does not have while simultaneously the object does not exist. The motivation for the super-illusion arises from the notion present in the literature where some philosophers claim that we are under the illusion of conscious experience as existing and simultaneously are under the illusion of the illusion of consciousness posing a unique epistemic challenge (Chrisley and Sloman 2016, p. 6). To further support the super-illusion as a viable account of the illusion of consciousness, it is necessary to be able to conceive of a particular functional implementation. It is the necessity of conceiving of an account of the illusion of consciousness in functional terms which I discuss in the next section.

4 The functional criterion

To establish the function criterion, I argue that a functional account of illusion must be conceivable for an account of the illusion of consciousness to be conceivable. Following the structure of necessity arguments introduced along with the definitional criterion (p. 7), the argument for the necessity of the functional criterion can be put as:

(1) A functional account of illusion is conceivable.
(2) An account of the illusion of consciousness is conceivable.
(3) If *(2)* cannot be true unless *(1)* is true then *(1)* is necessary for *(2)*.
(4) *(2)* cannot be true unless *(1)* is true.

(5) *(1)* is necessary for *(2)*.

Let us begin by considering the premise (4). The reason why one should believe the premise (4) to be true is based on the assumption that the illusion of consciousness is a type of illusion. I outlined the reasons for this assumption with the definitional criterion (p. 7). If the illusion of consciousness is not a type of illusion, then it is not an illusion. A functional account of illusion in the general case would, therefore, also include the illusion of consciousness. From this relationship one can infer that if one can conceive of a functional account of illusion the general case, then one can conceive of a functional account of the illusion of consciousness.

The proposition that the illusion of consciousness can be conceivably explained in functional terms is particularly important to the goals of illusionism. As Chalmers claims "the hard problem turns crucially on the claim that the concept of phenomenal consciousness is not a functional concept" (2018, p. 50). It is precisely the capacity of reducing conscious experiences to illusions that gives the capacity to the illusionist theory to explain away the hard problem. Frankish hints at the functional criterion with his claim that the illusion of consciousness must be accounted for in "broadly functional terms" (Frankish 2016, p. 14). The reasons for believing the claim that a functional account of the illusion of consciousness can explain away the hard problem can be explained using the easy and hard problem distinction (introduced on p. 4). The easy problems of consciousness can be explained in functional terms while it is not clear whether the hard problem of consciousness can be explained in functional terms. If the problem of conscious experience can be shown to be explainable via functionalism, then there is no hard problem of consciousness. Illusions are generally deemed to be explainable in functional terms. Thus, illusions are one of the easy problems. If conscious experience is a product an illusion, it is, therefore, explainable in functional terms and there is no hard problem of consciousness. This is why an account of the illusion of consciousness must be accounted for in functional terms.

In addition, the capacity of illusionism to explain away the hard problem can be also supported in the form of a debunking argument (Chalmers 2018, p. 44). According to the debunking argument concerning the hard problem, if there is a broadly reductionist explanation of our non-reductionist beliefs about consciousness, non-reductionist beliefs will not be justified (Chalmers

2018, p. 45). Chalmers lays out the argument in the following manner:

> (1) There is a correct explanation of our beliefs about consciousness that is independent of consciousness.
> (2) If there is a correct explanation of our beliefs about consciousness that is independent of consciousness, those beliefs are not justified.
>
> ---
>
> (3) Our beliefs about consciousness are not justified.

Since functionalism is a broadly reductionist theory and an account of the illusion of consciousness would explain our beliefs about the conscious experience, this argument can be said to hold for illusionism too. The explanation of our beliefs in the case of illusionism is the claim that our beliefs are based on the illusion of the existence of conscious experience. Since conscious experience is a product of an illusion and, thus, does not exist, our beliefs based on the existence of conscious experience not justified. In other words, if there is a functional explanation of the illusion of consciousness, our belief about the existence of the hard problem are not justified.

In sum, it is the capacity of functionalism to explain away the hard problem which establishes the premise (4) of the functional criterion argument (p. 13). This capacity can be either explained via the hard and easy problem of consciousness distinction or via a debunking argument. Nevertheless, one might ask what it is about functionalism that allows for the distinction between easy and the hard problem. This problem is also important for establishing the premise (1) of the functional criterion argument which requires the conceivability of a general functional account of illusions (p. 13). I confront these difficulties in the following section.

A suitable functional account

In this section, I outline a version of functionalism which could constitute a suitable foundation for a functional account of illusion. I do so by providing an overview of the historical variations of functionalism and providing an account of Chalmers' Combinatorial State Automaton (CSA).

A careful reader might be ask why would I speak of functionalism while the illusionist theory is supposed to maintain the consistency of physicalism. This distinction can be reconciled by the claim that functionalism is entailed in physicalism. Physicalism accounts for the world in terms of physical properties while functionalism describes the world in terms of functional roles. Since physicalism describes the world in terms of physical properties, it is dependent on these properties. For example, the dynamics of photons within a particular system does not translate to a description of oxygen particles, despite their possible commonalities. On the contrary, a functional description can translate between various physical systems. One could say that such a description is substrate independent. Thus, the advantage of a functional description of the illusion of consciousness is that it can translate between various physical systems, such as a brain or a computer, as long as they share the relevant functional roles. Thus, an instantiation of a functional concept can be said to be reducible to a physical concept.

The original formulation of computational functionalism by Putnam proposes that mental states and events are computational states of the brain, and so are defined in terms of "computational parameters plus relations to biologically characterized inputs and outputs" (1987, p. 7). What matters in such a computational account is a *functional organization* which is the way in which mental states are causally related to each other, to sensory inputs, and to motor outputs. Putnam proposed this framework as opposed to classical materialism—the hypothesis that mental states are brain states—and behaviorism—the hypothesis that mental states are behavioural dispositions (Place 1970, p. 44; Carnap 1959, p. 107). Computational functionalism is distinguished from other types of functionalism in that it explicates causal relations in terms of computational parameters (Shagrir 2005, p. 22). Computational functionalism has been since rejected as wildly implausible, even by Putnam himself.

There are other formulations of functionalism which are likely more suitable for the account of the illusion of consciousness. One of the prominent reformulations of computational functionalism is by Chalmers (1996, p. 309). He points two major shortcomings of Putnam's model: (1) the incapacity of the model to exhibit stable transitions in between states; and (2) the possibility for unexhibited stable transitions. These conditions cause the possibility that every ordinary open system could realize every abstract finite automaton. By an automaton we mean a set of states with a predefined set of transitions between these states. A finite automaton might be, for example, a Turing machine. The states of the machine can then be mapped on states of an ordinary open system, such as a brain or a rock. However, Chalmers argues that such a property of Putnam's theory undermines the use of the theory, unless one is willing to accept that every physical object can have a mind (1996, p. 310). Chalmers addresses these shortcomings by proposing the Combinatorial State Automaton (CSA)(1996, p. 325). Using the construction of the CSA, he shows that there can be a causal organization of a physical system mirroring the formal organization of an automaton while the automaton must employ constraints which is more likely to hold for brains rather than rocks (Chalmers 1996a, p. 332). In this way, he establishes a bridge between a computational theory and physical systems in such a way that the causal organization of physical systems mirrors the formal organization of an automaton. He defines implementation as follows:

> A physical system implements a given CSA if there is a decomposition of its internal states into substates $[s^1, s^2, ..., s^n]$, and a mapping f from these substates onto corresponding substates S^J of the CSA, along with similar mappings for inputs and outputs, such that: for every formal state transition $([I^1, ..., I^k], [S^1, ..., S^n])$ $\to ([S'^1 ...., S'^n], [O^1, ..., O^l])$ of the CSA, if the system is in internal state $[s^1, ..., s^n]$ and receiving input $[i^l, ..., i^n]$ such that the physical states and inputs map to the formal states and inputs, this causes it to enter an internal state and produce an output that map appropriately to the required formal state and output (Chalmers 1996a, p. 325).

If we consider the functionalist account by Chalmers in the context of the illusion problem, our toolkit consists of a formalism which describes the change in systems with the use of causal relations and their change based on the input and output relations with a particular set of predefined state transitions. In other words, the CSA allows for a step-by-step description of physical systems that is not constrained to a specific physical systems, but can explain a particular functional organisation across various functional system, including illusions. Among other things, the generalisability of functionalist makes an explanation in functional terms reproducible and testable. Generalisability, reproducibility, and testability can be said to be among the main properties which make a functional explanation superior to other forms of explanations.

It remains a subject of future research whether Chalmers' account of functionalism has the capacity to account for illusions in the general case, as well as the illusion of consciousness. Nonetheless, by narrowing down a specific account of functionalism, I hope to have shown the means by which a general account of illusions should be provided and establish it as conceivable.

5 The framework criterion

To establish the framework criterion, I argue that the illusion of consciousness is conceivable only if it is grounded in a general framework of the mind. Following the structure of necessity arguments introduced along with the definitional criterion (p. 7) and the functional criterion (p. 13), the argument for the necessity of the framework criterion can be put as:

> (1) It is conceivable for illusion of consciousness can be grounded in a general framework of the mind.
> (2) An account of the illusion of consciousness is conceivable.
> (3) If *(2)* cannot be true unless *(1)* is true then *(1)* is necessary for *(2)*.
> (4) *(2)* cannot be true unless *(1)* is true.
> ―――――――――
> (5) *(1)* is necessary for *(2)*.

Let us begin by considering the premise (4). The truth of this premise is contingent on whether we can conceive of an account of the illusion of consciousness without grounding it in a general framework of the mind. The reasons for accepting this premise lie in the question of whether one could conceive of an account of the illusion of consciousness independently of a general framework of the mind. Such a possibility appears absurd, as the illusion of consciousness is result of the operation of the faculty of the mind; thus, it should be consistent with a particular theory that explains its functioning.

Likely the most prominent and most relevant conceptual division is between the representationalist and anti-representationalist framework of the mind. The current illusionist discourse lacks a clarification of the particular commitments of the theory, although, given the functional criterion, it appears more likely to adopted as a representational theory. In general, the representational account postulates the existence of mental representations. Mental representations share some traits with other kinds of representations.

Simple examples of public and social representations include instances of writing or map-making. There is an extensive debate about the specific properties of mental representations and their viability as an explanans of the mind. I do not aim here to settle the discussion. However, I attempt to provide one avenue for illusionism to position itself within the representational theories of the mind.

A conceivable framework

There are existing representational frameworks which could constitute a foundation for the illusionist theory of consciousness. One of the perhaps most prominent frameworks is Graziano's 'attention schema.' Graziano proposes that the illusion of consciousness can be explained as an internal model, or representation, of attention. The model contains a notion of "a mechanistic method of handling data in which some signals are enhanced at the expense of other signals and are more deeply processed" (Graziano 2016, p. 98). Graziano's attention schema is an example of a theory with a plausible functional implementation. It accounts for illusions in a form of a mistake in an internal sensory model. Graziano says "an internal model, . . . with or without illusions, is an efficient, useful compression of data" (2016, p. 113). Nevertheless, the model never represents the world as it is. The specifics of Graziano's model are not as important, as its effects. By providing a plausible architecture for cognitive agents, he provides a framework to account for consciousness in functional terms. In turn, he provides a conceivable account of some of the processes which could give rise to the illusion of consciousness. Nevertheless, his approach is not consistent with super-illusion which I proposed earlier, as he utilizes attention as the core principle of his account.

Based on my account of the super-illusion, and similarly to the Graziano schema, I propose a speculative framework that could be responsible for the generation of the super-illusion, including the illusion of consciousness. The architecture mirrors the basic structure of Chalmers' CSA, in such a way that it contains an input and output layer and its relevant causal organization could be mapped on a physical system. Its internal states can be conceptualized as divided into a set of procedural layers. First, there is a sensory input layer, including all possible sensory modalities. Second, the agent employs strategies for object identification (OI) based on the relative differences in sensory input. Third, the agent employs an object manipulation (OM) layer which allows for a recognition of relationships between objects recognised by the object recognition layer. Finally, the agent uses the set of objects it recognises to construct an ontology (OC). The top three layers (OI, OM, & OC) can 'correct' each other in both directions ($\leftrightarrow$) based on relevant incongruities in its causal structure. Each one of these layers, given its causal structure, comprises a representational system. By that I mean that it represents the relevant causal relationships of the external world discerned from its inputs.

The super-illusion arises from corresponding mistakes (e.g. I^E or I^P) within the representational properties of our cognitive system. If one considers the definitions of the super-illusion, one can see that they are primarily concerned with the existence or properties of objects. Within philosophy, existence of objects and their properties is the subject of metaphysics. Thus, one can think of agents susceptible to the super-illusion as those who are engaged in the construction of metaphysical systems. The core procedure can be visualised as follows:

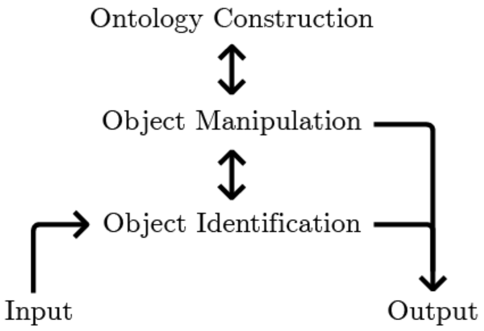

Figure 1: Super-illusion architecture

The super-illusion likely construes our understanding of consciousness but also of other concepts, such as time or free will. Smart argues that our understanding of the world is composed of a series of metaphysical illusions. He uses the notion of 'metaphysical illusion' as an assertion of a false ontology, but also more specifically as a result of conceptual or logical confusion, as well as the incorporation of psychological pressure in the false conclusion (Smart 2006, p. 175). Although he does not specifically argue for the illusion of consciousness, he provides an argument for the existence of the illusion of time, free will, and their connection to Armstrong's illusion of the truth of anti-materialism. In short, one might say that we are ontology constructing machines: navigating the world based on a class of objects and their relations. However, the processes constructing our ontologies sometimes go wrong—which is what gives rise to the super-illusion.

6 The empirical criterion

To establish the empirical criterion, I argue that the illusion of consciousness is conceivable if it is grounded empirical evidence concerning illusions. Following the structure of necessity arguments introduced along with the definitional criterion (p. 7), the functional criterion (p. 13), and the framework criterion (p. 18), the argument for the necessity of the empirical criterion can be put as:

> (1) An account of the illusion of consciousness can be supported by empirical evidence concerning illusions.
> (2) An account of the illusion of consciousness is conceivable.
> (3) If *(2)* cannot be true unless *(1)* is true then *(1)* is necessary for *(2)*.
> (4) *(2)* cannot be true unless *(1)* is true.
>
> ---
>
> (5) *(1)* is necessary for *(2)*.

Let us begin by considering the premise (4). The premise rests on the proposition that there one can only conceive of an account of the illusion of consciousness if the account can be supported by empirical evidence concerning illusions. For an account of the illusion of consciousness to be conceivable, it needs to establish commonalities with the functional accounts of other illusions. This proposition is coordinate with the definitional criterion, functional criterion, as well as the framework criterion. In theory, all illusions are easy problems—meaning that they can be explained in functional terms. Therefore, a functional explanation of the illusion of consciousness can be informed via the functional accounts of other illusions. As hinted by the name of this criterion, this undertaking relies heavily on empirical evidence from fields such as neuroscience, cognitive science, or computer science.

Illusions

Some of the most prominent examples of illusions are likely perceptual illusions. Examples of such illusions can span across several sensory modalities, but most familiar examples are likely visual illusions. Let us consider the Müller-Lyer illusion as an example of a perceptual illusion. Its two lines appear to have different relative length, despite being of identical length. In this example, we are perceiving an object which appears to have different geometrical properties than it does in reality. However, upon a closer examination, we can see that the lines are indeed the same.

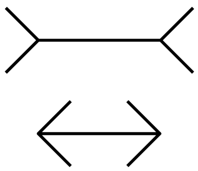

Figure 2: The Müller-Lyer illusion

There are various perceptual illusions which can be accounted for in functional terms. Some of these include the illusion of colour, which, according to Hall, can be encoded as dummy properties (2007, p. 201). Dummy properties are experiential properties which seem instantiated, but no physical objects that we might perceive instantiate them. They are arbitrary. For example, the illusion of colour refers to the fact that we perceive there to be colour, although there is no colour in the environment. Chalmers claims that we could come up with a computational system of colour representation which introduces a representational system that encodes qualities along with an R–G axis, a B–Y axis, and a brightness axis (2018, p. 26). In such way, a functional system could be said to implement the illusion of colour. Functional accounts of other illusions could possibly inform and perhaps even converge to provide a functional account of the illusion of consciousness.

In the case of the Müller-Lyer illusion, a closer examination does not make the illusion necessarily disappear, but we can recognize its illusory character. This is claimed to be distinct from the illusion of consciousness (Kammerer 2019, p. 6). When confronted with the proposition that conscious experience might be an illusion, in contrast to perceptual illusions, it appears difficult to recognise the illusory character of the illusion of consciousness.

This is hypothesised to be the case because the illusion of consciousness is an introspective illusion. According to Frankish, having the kind of inner life we have “consists of having a form of introspective representational mechanism that creates the illusion of a rich phenomenology” (2016, p. 22). An introspective representation does not concern the content of our perceptual input, but rather the content of our internal states. Chalmers claims that introspection “is an obvious place to start,” as “any intelligent system will need representations of its own internal states” (Chalmers 2018, p. 20). There is no reason to suggest that such a model of internal states could not be possibly accounted for in functional terms. One might ask when illusions arise in introspective systems and the answer likely is, they arise when the introspective systems get something wrong.

A good example of a mistaken introspective representation is the example of the “headless woman” (Armstrong 1968, p. 48). The “headless woman” argument is provided in defence of reductive materialism, arguably as one of the precursors to illusionism. The central thesis of the argument is that experience involves an illusion of the truth of anti-materialism. In his argument, Armstrong appeals to an example of a theatrical performance in which a woman is seated on a chair in front of a black background. Her head is covered with a piece of black cloth in such a way that she might appear headless. Armstrong advances the claim that we have the tendency to arrive at a false conclusion about what we perceive. In the case of the “headless woman,” an observer might conclude that the woman has no head instead of concluding that the observer only perceives the woman to have no head. Armstrong claims that this example constitutes an analog that is structurally equivalent to the illusion of the truth of anti-materialism. In other words, an observer might falsely leap from “I do not perceive that X is Y” (e.g. “I do not perceive that the woman has a head”) to “I perceive that X is not Y” (e.g. “I do perceive that the woman has no head”). Analogously, according to Armstrong, humans have the tendency to pass from “I am not introspectively aware that my experiences are mental states” to “I am introspectively aware that my experiences are not mental states.” In his account, this tendency creates the illusion of the truth of anti-materialism. As the absence of awareness of X is not equivalent to the awareness of the absence of X, Armstrong concludes that to keep materialism consistent, one can attribute such a false equivalence to an illusion.

The “headless woman” argument exemplifies an instance in which our introspective mechanism might fail and produce an illusion of anti-materialism.

There are likely many other examples of introspective illusions. Based on their commonalities one could perhaps suggest specific functional account of the illusion of consciousness. Nevertheless, this remains a project for further empirical work.

7 Conclusion

Illusionism features a seemingly absurd solution to an absurd problem—namely, by claiming that our conscious experiences can be understood by claiming that they do not exist. Although illusions appear implausible, there still remain various strategies that can address the inconceivability of the illusion of consciousness itself. I have argued that it is by clarifying the definitional criterion, the functional criterion, the framework criterion, and the empirical criterion that the illusion of consciousness might become conceivable. I have proposed possible strategies for each one of these criteria, including the reconsideration of the underlying concept in illusionism—an illusion—as a super-illusion. I have argued that there are distinct components of the super-illusion—the existence and the property illusion—which should be considered when accounting for the illusion of consciousness. There is still more work to be done in terms of the possibility of a functional account of the illusion of consciousness. However, I hope to have highlighted the prominence of the assumption that an illusion can be accounted for in functional terms and to have outlined some of the reasons for its further examination. At the very least, I hope to have provoked some thoughts about the nature of what we might think we understand the best: our experiences.[3]

[3] Thanks to Darcy Otto, Valerie Fowles, Daniel Lee, and Daniel Herrmann for discussion and comments.

Bibliography

Armstrong, D. M. 1968. "The Headless Woman Illusion and the Defence of Materialism." *Analysis* 29 (2): 48–49. ISSN: 0003-2638, accessed September 29, 2018. doi:`10.2307/3327124`. `http://www.jstor.org/stable/3327124`.

Carnap, R. 1959. "Psychology in Physical Language." In *Logical Positivism,* edited by A. J. Ayer. Free Press.

Chalmers, David. 1995. "Facing up to the problem of consciousness." *Journal of consciousness studies* 2 (3): 200–219.

———. 1996a. "Does a rock implement every finite-state automaton?" *Synthese* 108, no. 3 (September 1): 309–333. ISSN: 1573-0964, accessed March 9, 2019. doi:`10.1007/BF00413692`. `https://doi.org/10.1007/BF00413692`.

———. 1996b. *The Conscious Mind: In Search of a Fundamental Theory.* Google-Books-ID: oVqsjJvWgkMC. Oxford University Press, May 9. ISBN: 978-0-19-802653-2.

———. 2002. "Does conceivability entail possibility?" *Conceivability and possibility:* 145–200.

———. 2018. "The Meta-Problem of Consciousness." *Journal of Consciousness Studies* 25, no. 9 (January 1): 6–61. `https://www.ingentaconnect.com/content/imp/jcs/2018/00000025/f0020009/art00001`.

Chrisley, Ron, and Aaron Sloman. 2016. "Functionalism, revisionism, and qualia." *APA Newsletter on Philosophy and Computers* 16 (December 5): 2–13. ISSN: 2155-9708, accessed August 20, 2018. `http://www.apaonline.org/?computers_newsletter`.

Dennett, Daniel C. 1988. "Quining qualia." In *Consciousness in modern science.* Oxford University Press.

Fish, William. 2009. *Perception, hallucination, and illusion.* OUP USA. ISBN: 0-19-538134-3.

Frankish, Keith. 2016. "Illusionism as a theory of consciousness." *Journal of Consciousness Studies* 23 (11): 11–39.

Graziano, Michael SA. 2016. "Consciousness engineered." *Journal of Consciousness Studies* 23 (11): 98–115.

Hall, Richard J. 2007. "Phenomenal Properties as Dummy Properties." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 135 (2): 199–223. ISSN: 0031-8116, accessed March 9, 2019. `https://www.jstor.org/stable/40208747`.

Jackson, Frank. 1982. "Epiphenomenal Qualia." *The Philosophical Quarterly (1950-)* 32 (127): 127–136. ISSN: 0031-8094, accessed March 16, 2019. doi:`10.2307/2960077`. `https://www.jstor.org/stable/2960077`.

Kammerer, François. 2016. "The Hardest Aspect of the Illusion Problem–and How to Solve it." *Journal of Consciousness Studies* 23 (11): 124–139.

———. 2018. "Can you believe it? Illusionism and the illusion meta-problem." *Philosophical Psychology* 31, no. 1 (January 2): 44–67. ISSN: 0951-5089, 1465-394X, accessed December 27, 2018. doi:`10.1080/09515089.2017.1388361`. `https://www.tandfonline.com/doi/full/10.1080/09515089.2017.1388361`.

———. 2019. "The illusion of conscious experience." *Synthese* (January 2). ISSN: 0039-7857, 1573-0964, accessed March 4, 2019. doi:`10.1007/s11229-018-02071-y`. `http://link.springer.com/10.1007/s11229-018-02071-y`.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *The Philosophical Review* 83 (4): 435–450. ISSN: 0031-8108, accessed March 1, 2019. doi:`10.2307/2183914`. `https://www.jstor.org/stable/2183914`.

Place, Ullin T. 1970. "Is consciousness a brain process?" In *The mind-brain identity theory,* 42–51. Springer.

Putnam, Hilary. 1987. *Representation and Reality.* MIT Press.

Rescher, Nicholas. 2005. *Epistemic logic: a survey of the logic of knowledge.* University of Pittsburgh Pre. ISBN: 0-8229-7092-9.

Shagrir, Oron. 2005. "The Rise and Fall of Computational Functionalism." In *Hilary Putnam (Contemporary Philosophy in Focus),* edited by Yemima Ben-Menahem. Cambridge University Press.

Smart, J. J.C. 2006. “Metaphysical illusions.” *Australasian Journal of Philosophy* 84, no. 2 (June): 167–175. ISSN: 0004-8402, 1471-6828, accessed September 30, 2018. doi:10.1080/00048400600758912. http://www.tandfonline.com/doi/abs/10.1080/00048400600758912.

Turing, Alan M. 1937. “On computable numbers, with an application to the Entscheidungsproblem.” *Proceedings of the London mathematical society* 2 (1): 230–265.